主题爬虫技术研究综述 *

潘晓英 [1,2]，陈 柳 [1]，余慧敏 [1]，赵逸喆 [1]，肖康泞 [1]

(西安邮电大学 a.计算机学院; b 陕西省网络数据智能处理重点实验室，西安 710121)

摘 要：随着移动互联网的普及，网络信息指数增长，如何有效地提取和利用这些信息面临巨大挑战。首先介绍了主题爬虫的工作原理、分类；然后回顾了近年来国内外关于主题爬虫的研究状况，分析了各种主题相似度的方法以及搜索策略，得出相比于普通的爬虫系统基于网页内容和基于链接分析的爬虫系统，查准率、查全率都大幅度的提升；最后分析比较了主题网络爬虫两种动态搜索策略及未来研究方向。

关键词：网络爬虫；主题爬虫；相似度；网页内容；链接分析

中图分类号：TP393 doi: 10.19734/j.issn.1001-3695.2018.11.0790

Survey on research of themed crawling technique

Pan Xiaoying[1,2], Chen Liu[1], Yu Huimin[1], Zhao Yizhe[1], Xiao Kangning[1]

(a. *School of Computer Science & Technoligy*, b. *Shaanxi Key Laboratory of Network Data Intelligent Processing*, *Xi'an University of Posts & Telecommunications*, *Xi'an Shaanxi* 710121, *China*)

Abstract: : With the popularity of the mobile Internet and the growth of the network information index , how to effectively extract and utilize this information faces enormous challenges. Firstly, it introduced the working principle and classification of the topic crawler. Then it reviewed the research status of the topic crawler at home and abroad in recent years, analyzed the methods of similarity of various topics and the search strategy, and drawed the Web content based on the common crawler system. And the crawler system based on link analysis, it greatly improved the precision and recall rate. Finally, it analyzed and compared the two dynamic search strategies and future research directions of the topic Web crawler.

Key words: Web crawler; focused- crawler; similarity; Web page content; link analysis

0 引言

互联网是一个庞大的数据集合，网络信息资源产生的速度呈指数增加，如何有效地根据用户查询将数据分为相关和不相关数据，并利用这些信息是科研人员现如今面临的巨大的挑战。日常人们使用的检索工具有 Firefox、Google 等，但只提供粗略检索结果的传统搜索引擎，无法满足现在人类搜索的需求，提供精准的检索信息。为了弥补通用搜索引擎的缺陷，能够定向获取信息的检索工具——垂直搜索引擎出现。主题爬虫作为垂直搜索引擎的核心部分，如何使爬虫更精准、更快速的抓取信息，成为爬虫领域中的一个重要研究方向，引起了国内外众多研究人员的广泛关注。

本文介绍了爬虫工作原理、分类、系统结构、爬虫的关键技术，详细分析了基于网页内容的主题爬虫和基于链接结构分析的主题爬虫。实验结果表明与普通的爬虫系统相比，主题爬虫的查准率、查全率都有大幅度的提升。

1 网络爬虫的工作原理

网络爬虫，也称蜘蛛[1]。可以自动化浏览网络中的信息。搜索引擎离不开网络爬虫，网络爬虫的主要作用是在海量的互联网信息中进行爬取，抓取有效信息并存储。

图 1 为网络爬虫的实现原理及过程示意图。其中，初始的 URL 地址可以由用户人为地指定，也可以由用户指定的某个或某几个初始爬取网页决定。以初始 URL 开始，即种子 URL，当爬虫访问整个网页时，它会自动识别网页中所有 URL，并将其添加到待爬取 URL，按照一定的搜索策略访问待爬取 URL，采集对应 URL 的网页后将网页存储到数据库中，根据新的 URL 爬取网页，同时从新网页中获取 URL。重复上述的爬取过程。当爬虫符合整个系统设置的停止条件，则网络爬虫停止网页抓取。

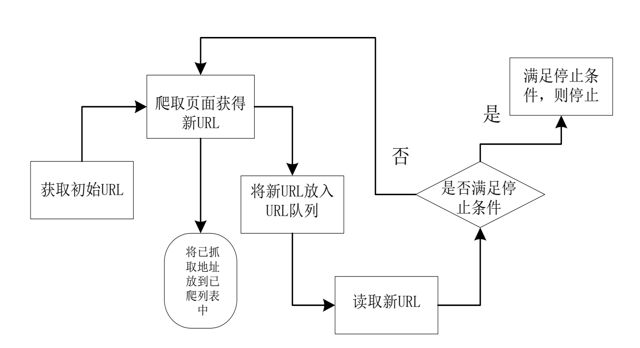


图 1 网络爬虫的实现原理及过程

Fig. 1 The realization principle and process of web crawler

2 网络爬虫的分类

网络爬虫按照实现的技术和系统可以分为通用网络爬虫（general purpose Web crawler）、主题网络爬虫（topical

收稿日期：2018-11-14；修回日期：2018-12-27 基金项目：国家自然科学基金资助项目（61373116）

作者简介：潘晓英（1981-），女，浙江丽水人，副教授，博士，CCF 会员，主要研究方向为人工智能、数据挖掘；陈柳（1993-），女，陕西西安人，硕士，主要研究方向为人工智能、主题爬虫，（chenliu@vdail.com）；余慧敏（1995-），女，陕西商洛人，硕士研究生，主要研究方向为人工智能，数据挖掘；赵逸喆（1994-），女，陕西榆林人，硕士研究生，主要研究方向为人工智能，数据挖掘；肖康泞（1997-），男，陕西西安，本科，主要研究方向为人工智能，数据挖掘.

crawler）、增量式网络爬虫（incremental Web crawler）、深层网络爬虫（deep web crawler）。

通用网络爬虫又叫做全网爬虫，其爬取目标在整个互联网中。由种子 URL 开始，爬虫系统开始访问网页，采集网页所有超链接。为了防止获取重复的 URL，将爬取到的网页信息存储在原始数据库中，然后对网页进行解析，并根据网页搜索策略爬取新的 URL。重复上述过程，直到爬取到的 URL 符合停止条件，则完成整个爬虫过程。这种面向全网的检索工具，无法准确提供用户特定的需求[2]。因此，提出了面向特定主题需求的网络爬虫：主题网络爬虫，它比通用网络爬虫多出几步，即目标的定义、无关链接的过滤、下一步爬取 URL 地址的选取。

主题网络爬虫可以按照对应的主题有目的地进行爬取，聚焦网络爬虫将目标定位在互联网中与主题相关的页面中，初始 URL 的获取是通过对抓取目标的定义以及相关的描述。为了帮助爬虫更有效的发现与主题相关的 URL，需要对主题准确的描述，然后解析网页内 URL，判断网页与主题的相关度，根据网页搜索策略预测链接的主题相关度并确定 URL 优先级。在聚焦网络爬虫中，不同的爬取顺序会导致爬虫的执行效率不同，因此需要依据搜索策略来确定下一步需要爬取的 URL 地址并存储。整个主题爬虫不断重复上述过程，当符合爬虫系统中规定的停止条件，则停止爬取过程。

3 网络爬虫的系统结构

网络爬虫系统分为网页获取、网页过滤以网页存储三大模块。主题爬虫为了定向的抓取有效信息，对三大模块进行适当修改并增加了网页分析模块用于计算网页相似度，如图 2 所示。主题网络爬虫的关键是确定主题并对主题进行详细描述，在系统抓取页面之前给定网页文本与主题的相关性，使爬虫系统尽可能多地筛选出和主题相关页面，减少无关页面，从而使主题爬虫返回的结果具有较高的准确率。相比较通用爬虫，主题爬虫优势有如下几点[1]：a)相比通用爬虫只能提供粗略的信息，主题爬虫主题明确且系统能够精准地获取有效信息；b)主题爬虫在存储网页 URL 需要判断该 URL 与主题的相关性，尽可能筛选出与主题相关的页面。

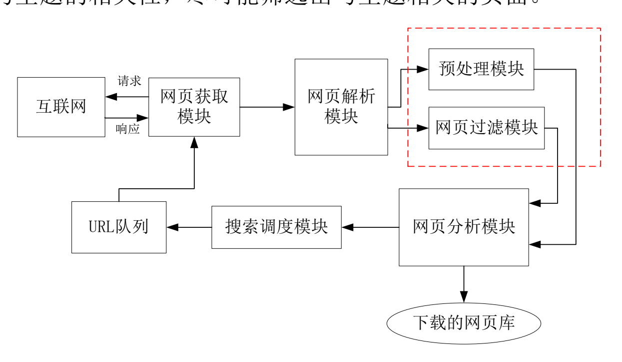


图 2 网络爬虫的系统结构

Fig. 2 System structure of Web crawler

爬虫系统主要模块介绍如下：

a) URL 队列。URL 队列主要用来存放各种超链接，如系统未爬取的网页链接，即待爬取 URL 队列；随着爬虫系统运行更多的链接被爬取，为避免爬虫系统爬取相同页面，已爬取的链接存放已爬取队列；未完成下载的链接被存放在错误队列。

b) 网页获取模块[3]。网页获取中需要模拟客户端发送 HTTP 请求，获取服务器端的响应后下载网页，完成爬虫系统爬取工作。同时，爬虫系统为了确保整个网络爬虫的正常工作和效率，防止抓取同一网页，在网页获取模块中设定超时机制，超过一定抓取时间的网页将被舍弃。

c) 网页解析模块。网页解析模块是衔接其他模块的中枢，是整个爬虫系统主要的部分。该模块提取采集的 HTML 形式网页中的重要信息链接、文本等，同时利用获取的内容信息，为后期网页的主题相关度计算做铺垫。

d) 网页过滤模块。该模块用来筛选与主题有关的 URL，通过筛选抓取与主题相关的页面，确保主题爬虫系统的准确率。

e) 搜索调度模块。为确保爬虫对 URL 更有效、合理地访问，网络爬虫会根据网页制定合理的搜索规则。常见的网页搜取策略分为深度优先、广度优先和最佳优先三种。由于深度优先存在一定问题，最常用的是广度优先和最佳优先两种搜索方法。

f) 网页存储模块。网页存储模块将网页解析模块解析出来的数据通过文件或数据库的形式存储起来，从而为搜索引擎完成检索功能做好准备。

g) 预处理模块。该模块是将网页解析模块获取的网页内容等信息进行处理，通过对文本的分词、去停用词、词干化等预处理，将文本内容转换为计算机能够识别的数学模型，为后期主题网络爬虫中网页分析模块进行主题相似度计算做准备。

h) 网页分析模块[4]。该模块是主题爬虫的核心部分，网页分析模块分为两部分，第一部分是主题相关度判断，用于判断网页的主题相关性；第二部分为主题相关度预测，预测网页 URL 与主题相关度，通过搜索策略，优先访问与主题相关的 URL。

4 网络爬虫关键的技术

4.1 网页获取

网络爬虫的基本原理是模拟浏览器进行 HTTP 请求，爬虫客户端通过 HTTP 请求向 Web 服务器发送请求，获取服务器端的响应后下载网页，完成爬虫系统爬取工作。

4.2 网页解析

网页解析主要是一个网页去噪的过程，互联网中以 HTML 为架构承载网页的各种信息。网页去噪主要是网页内容正文抽取。主题爬虫提取网页中的内容时，需要分析页面的 HTML 结构，从中提取页面的有效信息。常见的方法有通过 BeautifulSoup 对 HTML 结构解析、利用正则表达式抽取文本数据。

BeautifulSoup 主要是 Xpath 和 CssSelector 方法，针对网站的 HTML 标签可以提取出所需要的有效信息，可以选择 tag、id、class 等多种方式进行定位选择。Chrome、firefox 浏览器已经对页面的各个节点做好了标记，可以直接复制 Xpath 或者 CssSelector 使用，相比较正则表达式，BeautifulSoup 方便初学者使用。但结构复杂的页面中，BeautifulSoup 并不是一种高效的方法，在使用这种方式提取有效信息，就需要要求页面的结构固定，相同字段的 tag、id、class 都必须相同，所以在复杂的页面结构中，就需要采取正则表达式来提取有效信息，正则表达式比较复杂，需要花时间去研究；但是对于提取页面字符串结构的信息，处理速度很快，高效便捷。

4.3 数据存储

爬虫抓取后的数据，一般选择两种存储方式：本地保存 csv、excel 格式或者直接存储到数据库。对于量大的数据可以直接保存本地，对于数据量大的爬虫一般选择保存在数据库中，方便储存同时也方便后期进一步对数据的分析、处理

等。 用 python 写爬虫的过程，直接可以采用 python 中自带的 csv 包、新建 csv 或者 excel 格式的表格。设置边爬边存储 csv 或 excel 中写入数据库中分为两种形式，一种是关系型数据库 MySQL、SQLServer；一种是非关系型的数据库 mongodb、ssdb、hbase 等。

写入数据库有两种思路，一种是等所有的数据都爬完，集中一次向量化清洗，一次性入库；另一种是爬一次数据清洗一次就入库。对于大规模爬虫来说，稳定性是要考虑的重要因素，在长久的爬虫过程种，不可避免地出现一些网络错误，在这种情况下第一类爬出的数据会变成无用数据；而第二类则避免了类似问题，并且单次清洗和入库较快，对整体入库时间不会产生影响，故选择第二类方法作为写入数据库的方式。

4.4 主题判别

主题判别[5]的主要作用是判断爬取网页的主题相关性，第一步就是思考如何定义主题。主题判别的问题大多被当做一个文本分类的问题来探索。目前，研究人员结合网页中链接的锚文本、网页标签等来计算网页中 URL 与主题的相关度。因此主题相关性的计算也是不同主题爬虫的区别之处。常用的主题相似度判别算法有向量空间模型、语义相似度[3]。

1）向量空间模型

向量空间模型概念简单，将文本处理转换为在向量空间上的向量运算，将每一篇文档表示为向量空间上的某一维度，通过计算向量在空间的相似度来衡量文档之间的相似度。

2）语义相似度

汉语不同于英语，对某个事物的描述有多种不同的描述方式，尤其是近年来研究人员困惑的问题自然语言处理中语义理解，识别一段文本的含义，传统的分词、统计词频不能准确理解文本信息所表达的意思，降低文本含义识别的准确度。文本中能够观察到的量只有词频和文档频率两个，在文本语义的分析方法，是一种对以这两个量为主要思想的计算基础，使得计算机能够“懂”人类的语言。

4.5 网页搜索策略

主题爬虫是定向爬虫，具有特定的主题其目标就是快捷准确地完成与主题相关页面的搜索。网络搜索策略[1]主要目的就是使爬虫有次序、有目的地搜索，运用合理的搜索策略可以保证主题爬虫选择更合理的爬行路径，高效地完成网页爬取任务。

网络搜索策略依据搜索方式的不同分为静态搜索策略和动态搜索策略。静态搜索策略和动态搜索策略主要区别是有无事先确定搜索规则。静态搜索策略依照确定的规则进行搜索，搜索策略的规则不会因为网页结构、文本信息的改变而改变；动态搜索策略以高效、快速完成爬取任务为第一宗旨，实时调整搜索路线；互联网是由网页和超链接构成的一个整体，根据分析对象不同，动态搜索策略可分为基于文本内容的搜索和基于链接关系的搜索。

网页中不同的内容信息反映网页不同的含义，标题、关键词、文本内容等都是网页中最具有代表性的信息，主题爬虫获取网页后依据网页全局文本信息或网页局部信息计算主题相关度。动态搜索策略需要快速计算网页链接相关度，因此基于局部文字的搜索策略是主题爬虫较常用的一种搜索策略，该计算相关度该方法计算量小，能够在较短时间得到 URL 的主题相关度；基于网页全局文字的搜索策略利用网页所有文本信息耗时过长。基于文本内容的经典的搜索策略有 Fish-Search、Shark-Search。目前研究人员提出的基于链接分析的搜索策略都是建立以三条标准为基础：a)网页的引用，网页的价值与网页被引用成正比；b)网页之间存在被引用关系，则网页结构、内容信息相似度越大；c)结构信息合理的网页易被引用。基于链接分析的搜索策略利用网页中的链接来进行分析并预测网页主题，最后评估网页 URL 的优先级。目前经典的搜索策略有 Page Rank、HITS 和 Hill Top。

5 主题爬虫的研究方向

近几年来，研究者们为了提升主题爬虫获取页面时的准确度和高效性，通过在主题相似度和搜索策略上制定爬行策略和算法。目前国内外对主题爬虫的研究主要分为以下几个方向。

5.1 基于网页内容的主题爬虫

网页中不同的内容信息反映网页不同的含义，标题、关键词、文本内容等都是网页中最具有代表性的信息。王锦阳[1]利用这一特点提出利用标题构造精简内容子树来判断网页主题，利用语义相似性改变向量空间模型对主题的相关性进行判定，解决了传统向量空间模型缺乏在文本语义判定中的问题，提高了判断网页主题相关的识别率，主题爬虫采集信息的准确率大大提高。

周米雪[4]在主题爬虫的启发下设计面向医疗领域的垂直搜索引擎，在抓取网页后，分别从网页中的超链接、元信息、词库进行主题相关度判别，有效合理地筛选出与主题相关的页面；并针对传统的 PageRank 算法的不足，合理地引进时间反馈因子、权威性因子、主题相关度因子。实验结果表明，医疗垂直搜索引擎的查准率明显提高。

李宏志等人[5]构建了 KNN 分类器来判断网页之间的主题相关性，采用 IK Analyzer 实现网页内容的中文分词，通过 TF-IDF 算法实现网页内容的特征提取。实验结果表明，基于 KNN 分类的网络爬虫在区分网页主题时准确率会随着网页中文档数量的增加而升高，同时分类的效果、稳定性也优于传统的 PageRank 和 Bayes 算法效果。

张莉婧等人[6]将主题爬虫应用到图书主题上，设计了一种新的面向图书的主题爬虫算法 ODP2EVSM 到出一种面向图书主题的爬虫算法。该算法主要由两部分组成：为了准确、详细的描述主题作者首先采用基于开放式分类目录系统（ODP）进行关键词动态扩充的动态关键词扩充的主题描述方法；然后判断网页与主题是否相关采用基于词项语义扩展度的向量空间模型（VSM）主题相关度算法。

李辉等人[7]利用向量空间模型对主题爬虫算法中的内容相似度进行计算，爬虫在采集页面时有效地筛选出和主题相关度高的网页，同时提高了爬行效率和抓取的准确度。为验证该算法爬行的准确率，将该算法应用在养殖投入品质量信息监管系统，测试表明该系统运行稳定、采集信息准确度高。

姬祥[8]利用农产品价格样本得到一个 SVM 分类器,以 SVM 分类器的支持向量为训练样本构建一个 KNN 分类器,有效地对抓取到的页面进行分类。为了精准高效地收集所有农产品价格信息，在不同情况下分别采用 SVM 分类器和支持向量 KNN 分类器来保证抓取网页准确性。

网页文本信息中存在一词多义的问题，为了解决这一问题，孟竹[9]提出了使用语义模型表示并结合点对互信息，根据上下文判断该词在文中的含义，并联合网页链接判断是否与主题词相关的判定，筛选出与主题词相关的网页链接，并得到客观的实验结果；

Wang[10]以电子产品品牌等专业词汇，将主题词汇扩展到我的词典中，使其称为一个典型的专业词汇，在很大程度上提高了查询的准确性，通过改进开源爬虫框架 Heritrix 建立

了电子产品搜索引擎。研究表明，该下载方法可以满足电子产品搜索平台需求。

Song 等人[11]在介绍关键词和支持向量机模型的基础上，提出了一种动态主题爬虫系统，能够有效地获取目标信息。该方法可灵活应用于信息安全、企业公关危机管理等领域。

Dahiwale 等人[12]提出依据语义相关性来判断主题相关性，通过在下载页面之前使用 Meta 标签作为计算相关性的主要信息来源，预测语义相似性的基于语义的聚焦 Web 爬虫的方法，通过搜索分析发现爬虫过滤了大量不相关的网页链接，系统采集的文档质量也有所提高。

主题爬虫在主题相似度判别算法制定各种算法，整个过程涉及文本相似度的判断。目前，基于网页内容的主题爬虫计算文本相似度的判断方法大致分为两类，一类是基于字词统计模型，如向量空间模型（VSM）；另一类是基于语义理解模型。研究人员希望使用语义相关性使网络爬虫可以获得更精确的结果。整个主题相似度判别过程中，首先确定主题爬虫的主题，再根据网页内容、结构信息计算网页主题相关度和抓取 URL 的相关度，依据网页主题相关度判断待抓取链接和抓取链接的优先级。此类爬虫通常能获得较高的准确率。

表 1　基于网页内容的主题爬虫方法

Table 1　Topic crawling method based on Web content

方法	文献	查准率	召回率	F 值
基于改进 PageRank 算法	[4]	0.7	\	\
基于 KNN 分类算法的主题网络爬虫	[5]	0.75	\	\
基于 ODP 主题描述和 VSM 主题相关度改进	[6]	0.64	0.24	0.24
基于词向量语义模型构建主题爬虫	[9]	0.46	0.69	0.44
基于 SVM 分类器的支持向量构建 KNN 分类器	[8]	0.80	\	\
基于关键词和 SVM 的动态主题爬虫	[11]	0.92	\	\
基于 URL 和锚文本语义特征改进	[12]	0.69	\	\

判断主题爬虫抓取性能主要指标有查准率、召回率（查全率）、F 值三条。表 1 给出了目前研究人员提出基于网页内容的主题爬虫算法的部分实验数据。实验结果表明，网页内容详细反映了网页的主题信息等，基于网页内容的主题爬虫算法的改进很大程度地提高了爬虫系统查准率与查全率。

5.2　基于链接分析的主题爬虫

传统的基于网页内容评价的搜索策略往往会忽略网页间链接的相关性，基于链接分析的搜索策略忽略了网页正文内容，造成“主题漂移”的现象。蔡光波 [13]结合基于内容评价的 Fish-Search 算法和基于链接分析的 PageRank 算法从页面内容和页面间的链接关系两个方面进行考虑，将网页文本内容和网页链接结合使用、取长补短，从而计算出页面内容与主题间的相关性。爬虫系统结果验证表明，查准率明显提高。

胡萍瑞等人[14]依据网页中 URL 链接的结构、语义特征的相似性，而不同模块中 URL 链接特征差异较大，提出了基于 URL 模式集的主题爬虫，通过区分 URL 特征之间的差异来判断主题之间的相关性，并根据各模式的重要度预测待抓取 URL 的优先级，保证爬虫的查准率和查全率，提高爬虫效率。

张金等人[15]为确保获取的 URL 都是主题相关度高的页面，提出了基于页面子链接分析的链接排序算法，通过考虑子链接的相关度对当前的链接相关度进行加权，抓取过程中获得较高相关度链接，然后加权计算所得的得分对链接队列进行相关度高低的排序，从而提高了爬取的准确性。

史宝明等人[16]提出基于链接模型的相关性判别算法，利用计算待分析 URL 之间的主题相关度，先实现结果证明相比传统爬虫算法，提出的方法效率更高。

Liu 等人[17]通过采用 VIPS 算法分析网页的深度，在相关链接的预测中，采用多粒度鲨鱼搜索算法，同时结合基于查询的命中算法，改进了鲨鱼搜索算法爬行策略。新算法不但弥补了 Shark 和 HITS 两种算法的缺点，减少了噪声环节同时消除“主题漂移”现象。

网页中锚文本包含了页面中丰富的信息，Kumar 等人[18]依据这一特点建立页面分析器的组件被用来理解页面内容和页面中锚文本上下问的主题，页面分析器的输出用于进行爬行决策，从链接中提取信息，并引导爬虫在相关领域的特定爬行。

刘韶涛等人[19]通过结合基于内容的链接选择 Best-First 算法，引入能够体现链接价值的 HITS（hyperlink induced topic search）算法，将两种算法相结合，设计出新的链接选择策略。该算法将页面内容与链接结构融合起来考虑，有效地提高了爬虫在下载过程中的主题相关性和权威性。

Pant 等人[20]中提出爬行器在使用链接上下文的情况下自动导航 Web 的超链接结构，以预测相对于某些起始主题或主题的相应超链接的优势。使用由支持向量机引导的主题爬虫，研究了链接上下文的各种定义对爬行性能的影响。

Shen 等人[21]在结合 Web 内容分析，提出了一种基于复杂网络中局部社区的抓取方法。整个爬虫被分为两个部分，首先，利用社区发现算法对 Web 站点之间的链接结构进行分析，构建给定主题的网站；其次，对 Web 页面的所有主题相关分析和链接预测都在这个组内进行。

Gupta 等人[22]通过锚文本确定网页含义，提出标签树方法和解析方法提取链接上下文的方法。标记树方法将有助于找到锚文本的概念，并且该概念将由 LALR 解析器使用，用于提取链接上下文的算法。

Peng 等人[23]认为网页中锚文本不能有效地表达网页含义，可能会误导主题爬虫爬行方向，提出将网页划分为更小的区域来避免网页中高度相关区域被遮挡，并且根据划分区域的相关性选择使用链接上下文信息，以提高重点网页的采集。

Geng 等人[24]为了提高主题爬虫的采集效率和准确度，基于传统主题爬虫技术上提出了 HTML 分析和文本密度结合对网页文本提取，并考虑加入多因素计算相似度方法，即新闻文本和文本作为不同的参考因素。该方法明显提高了主题爬虫在网页文本的准确性。

Shark-Search 算法在距离相关页面集更近的距离内搜索时表现出良好的性能，但它缺乏“全局”。PageRank 算法是一种迭代算法，因此紧密相连区域中页面的权重必然会增加，从而导致“主题漂移”现象。Qiu 等人[25]将 Shark-Search 与 PageRank 算法合并，该算法分为两部分：采用 Shark-Search 算法计算网页得分，在用 PageRank 计算页面之间 URL 链接的权重值定义页面的重要性，同时弥补了两个传统算法的缺陷。结果表明该算法适用于大量页面的采集，以获取有效的网页信息。

互联网中数十亿的网页通过万维网上的超链接链接，研究人员试图通过有效的方式获取链接上下文的含义，从而对链接上下文的解析和提取，或者基于网页内容对传统链接选择算法改进，使网络爬虫采集过程中准确度提升。该类算法通过分析网页链接判断网页的重要性、强调了页面链接的权威性对用户的需求是有意义的，同时从网页正文、链接锚文本以及锚文本上下文网页内容分析和链接分析结合解决了主题漂浮问题，提高主题爬取的准确性。

表 2 给出了目前研究人员提出基于链接分析的主题爬虫

算法的部分实验数据。实验结果表明，基于网页链接分析的主题爬虫弥补了基于网页内容主题爬虫只考虑了页面内容的忽略了网页子链接形成对主题爬虫的影响的缺陷，且基于网页与链接同时研究会获取更精准的采集效果。

表 2 基于链接分析的主题爬虫方法

Table 2 Topic crawler method based on link analysis

方法	文献	查准率	召回率	F 值
基于 URL 模式集的主题爬虫	[14]	0.69	0.52	0.61
基于页面子链接分析的链接排序算法	[15]	0.55	\	\
VIPS 分析网页深度+多粒度鲨鱼搜索算法	[17]	0.66	\	\
分类器引导的主题爬虫且链接上下文	[20]	\	0.61	\
基于 Best-First 算法+HITS 算法	[19]	0.61	0.75	\
基于内容分块-选择性链接上下文的聚焦爬虫	[23]	\	0.80	\
HTML 分析+文本密度分析+多因子相似度	[24]	0.67	0.48	\

6 爬虫系统在各领域的应用

随着网络信息的指数增长，为创造更精准的检索工具，面向某一特定主题服务型垂直搜索引擎成为研究热点，因此不同领域的主题爬虫接踵而来。

智慧农林的兴起张露露 [27]设计了面向病害虫主题搜索引擎，构建领域主题词典对主题详细描述，同时考虑网站链接和网页内容设计满足该领域的主题搜索引擎；李辉等人[7]采用主题爬虫作为养殖投入品质量信息监管系统中对互联网中海量信息获取的关键步骤，有效避免了下载无关页面，提高信息采集的查准率、查全率；为了准确预测农产品价格涨幅，孟繁疆[28]等人构建农产品价格主题搜索引擎，该系统在收集农产品价格数据和价格变化的主要因素起到重要作用。

为助于整个人类健康，互联网技术在医疗领域逐渐扩展，尹曼 [29]通过分析医疗器械产品特点以及从业人员和消费人员不同的需求构建了面向医疗器械垂直搜索引擎；周米雪 [4]从主题相似度、PageRank 算法两方面作出改进构建医学垂直搜索引擎，经测试该搜索引擎查准率明显升高；李学博 [30]通过对互联网中存在的中医药信息分析设计中医药领域主题爬虫，该系统以最便捷、快速的方式从互联网中获取中医药信息，给人们提供可靠、精准的医疗健康信息服务。

刘灿等人[31]采用主题爬虫技术设计面向个性化推荐的教育新闻爬取，为人们能够准确及时获取教育类新闻；Hu R[32]等人提出基于谷歌的全栈技术 MEAN 开发了一种高效的定向爬虫（Mongo DB + Express + Angular JS + Node.js）堆栈和一个快速灵活的 Javascript 文档对象模型模块，称为 Cheer IO，在实际项目中该系统提供了大量有效数据；李翔宇 [33]设计开发了生物安全领域的主题爬虫，旨在从万维网海量信息中对该领域信息知识的精准获取。关卫国 [34]采用主题爬虫技术采集有关食品接触材料安全信息，这对食品接触材料安全领域网络舆情具有重要意义。

7 主题爬虫的发展趋势

目前为止，研究人员在主题网络爬虫上作出大量研究，但针对主题爬虫性能方面还有很大的研究空间，分为一下几点：

a) 网络爬虫都是固定的搜索策略，面对互联网中不同网站之间网页组织形式的不同，固定的搜索模式无法高效地抓取，如何通过集成爬取规则的方法来提高主题爬虫性能有待研究。

b) 宽泛的主题利用网页内容和链接上下文构建主题爬虫可以有效地计算出主题相关度，但针对较细化的主题存在一定局限性，例如对关键词描述不够准确，主题爬虫在采集信息时查准率、查群率都会降低，从语义角度改进对主题特征词的选取成为未来主题爬虫技术的研究热点。

c) 出于对网站信息的保护，设计网站时会出一套反爬虫策略来阻止爬虫抓取数据。针对反爬虫策略，研究人员引入分布式网络爬虫等高级爬虫来获取海量信息，但越高级的爬虫相应的开发成本高，能否设计出低成本的高级爬虫有待研究。

d) 网络舆情监控系统中对某热点话题信息采集，传统方法无法准确对主题进行准确描述，若利用热点话题具有的最显著特点：时间性，明确该话题产生时间、发展时间、消逝时间等。比如在食品安全主题的突发话题检测技术研究，对食品安全话题进行实时跟踪，增加话题时间变化度概念。

参考文献：

[1] 王锦阳. 主题网络爬虫的并行化研究 [D]. 成都: 西南石油大学, 2017. (Wang Jinyang. Parallelization of thematic Web crawlers [D]. Chengdu: Southwest Petroleum University, 2017.)

[2] 彭小明. 主题爬虫的设计与实现 [D]. 北京: 北京邮电大学, 2013. (Peng Xiaoming. Design and implementation of the theme crawler [D]. Beijing: Beijing University of Posts and Telecommunications, 2013.)

[3] 王聪睿. 主题爬虫关键技术研究 [D]. 石家庄: 石家庄铁道大学, 2015. (Wang Congrui. Research on key technologies of subject reptiles [D]. Shijiazhuang: Shijiazhuang Railway University, 2015.)

[4] 周米雪. 基于改进PageRank算法的医学垂直搜索引擎的研究与实现 [D]. 西安: 长安大学, 2017. (Zhou Mixue, Research and implementation of medical vertical search engine based on improved PageRank algorithm [D]. Xi’an: Chang’an University. 2017.)

[5] 李宏志, 宋婕. 基于 KNN 分类算法的主题网络爬虫 [J]. 宜宾学院学报, 2017, 17 (12): 61-65. (Li Hongzhi, Song Jie. Thematic Web crawler based on knn classification algorithm [J]. Journal of Yibin University, 2017, 17 (12): 61-65.)

[6] 张莉婧, 曾庆涛, 李业丽, 等. 面向图书主题的爬虫算法研究 [J]. 计算机科学, 2017, 44 (b11): 460-463. (Zhang Lizhen, Zeng Qingtao, Li Yeli, *et al*. Research on crawling algorithm for book theme [J]. Journal of Computer Science, 2017, 44 (b11): 460-463.)

[7] 李辉, 张标, 吴文良. 基于主题爬虫算法的养殖投入品质量信息监管系统 [J]. 江苏农业科学, 2017, 45 (8): 210-214. (Li Hui, Zhang Biao, Wu Wenliang. Quality information supervision system for aquaculture inputs based on subject reptile algorithm [J]. Jiangsu Agricultural Sciences, 2017, 45 (8): 210-214.)

[8] 姬祥. 农产品价格主题搜索引擎的研究与实现 [D]. 哈尔滨: 东北农业大学, 2017. (Ji Xiang. Research and implementation of agricultural product price theme search engine [D]. Harbin: Northeast Agricultural University, 2017.)

[9] 孟竹. 词向量语义模型研究及在主题爬虫系统中的应用 [D]. 北京: 中国地质大学, 2017. (Meng Zhu. Research on semantic model of word vector and its application in subject reptile system [D]. Beijing: China University of Geosciences, 2017.)

[10] Wang Aihua. Design and implementation of vertical search platform for electronic product information [C]// Proc of International Conference on Robots & Intelligent System. Washington DC: IEEE Computer Society, 2017: 101-104.

[11] Song Biao, Zhu Jianming, Zhang Jianguang. A research of dynamic theme crawler based on keywords and support vector machine [C]// Proc of International Conference on Management Science &

Engineering. San Francisco: IEEE Press, 2014: 21-26.

[12] Dahiwale P, Raghuwanshi M M, Malik L. Design of improved focused Web crawler by analyzing semantic nature of URL and anchor text [C]// Proc of International Conference on Industrial and Information Systems. San Francisco: IEEE Press, 2015: 1-6.

[13] 蔡光波. 面向主题的多线程网络爬虫的设计与实现 [D]. 兰州：西北民族大学, 2017. (Cai Guangbo. Design and implementation of topic-oriented multi-threaded web crawler [D]. Lanzhou: Northwest University for Nationalities, 2017.)

[14] 胡萍瑞, 李石君. 基于 URL 模式集的主题爬虫 [J]. 计算机应用研究, 2018, 35 (3): 694-726. (Hu Pingrui, Li Shijun. Theme crawler based on URL pattern set [J]. Journal of Computer Applications, 2018, 35 (3): 694-726.)

[15] 张金, 倪晓军. 基于语义树与 VSM 的主题爬取策略研究 [J]. 计算机技术与发展, 2017, 27 (11): 66-70. (Zhang Jin, Ni Xiaojun. Research on topic crawling strategy based on semantic tree and VSM [J]. Computer Technology and Development, 2017, 27 (11): 66-70.)

[16] 史宝明, 贺元香, 吴崇正. 主题搜索引擎中爬虫搜索策略的研究 [J]. 计算机工程与应用, 2014, 50 (2): 116-119. (Shi Baoming, He Yuanxiang, Wu Chongzheng. Research on crawler search strategy in topic search engine [J]. Computer Engineering and Applications, 2014, 50 (2): 116-119.)

[17] Liu Naiwen, Yao Rongbao. The crawling strategy of shark-search algorithm based on multi granularity [C]// Proc of International Symposium on Computational Intelligence and Design. San Francisco: IEEE Press, 2016: 41-44.

[18] Kumar N, Singh M. Framework for distributed semantic Web crawler [C]// Proc of International Conference on Computational Intelligence and Communication Networks. San Francisco: IEEE Press, 2016: 1403-1407.

[19] 刘韶涛, 李洪胜. 融合链接结构的主题爬虫算法 [J]. 华侨大学学报: 自然科学版 , 2017, 38 (2): 195-200. (Liu Yutao, Li Hongsheng. The theme reptile algorithm based on fusion link structure [J]. Journal of Huaqiao University :Natural Science , 2017, 38 (2): 195-200.)

[20] Pant G, Srinivasan P. Link contexts in classifier-guided topical crawlers [J]. IEEE Trans on Knowledge & Data Engineering, 2005, 18 (1): 107-122.

[21] Shen Guilan, Sun Jie, Yang Xiaoping. A focused crawling method based on detecting communities in complex networks [J]. Journal of Henan Normal University, 2014, 9 (8): 187-196.

[22] Gupta S, Yadav S. Extraction of link context using tag tree and LALR parsing [C]// Proc of Information & Communication Technologies. San Francisco: IEEE Press, 2013: 253-257.

[23] Peng Tao, Liu Lu. Focused crawling enhanced by CBP-SLC [J]. Knowledge-Based Systems, 2013, 51 (1): 15-26.

[24] Geng Zhingqiang, Shang Dirui, Zhu Qunxiong, *et al.* Research on improved focused crawler and its application in food safety public opinion analysis [C]// Proc of Beijing, Chinese Automation Congress. 2017: 2847-2852.

[25] Qiu Lei, Lou Yuansheng, Chang Ming. Research on theme crawler based on shark-search and PageRank algorithm [C]// Proc of International Conference on Cloud Computing and Intelligence Systems. San Francisco: IEEE Press, 2016: 268-271.

[26] Xiao Jiang, Ji Jie. The application of focused crawler based on Heritrix in internet public opinion system [J]. Electronic Design Engineering, 2015, 23 (6): 29-31.

[27] 张露露. 基于分布式采集策略的病虫害主题搜索引擎研究 [D]. 哈尔滨：东北林业大学, 2017. (Zhang Lulu. Research on pest and disease subject search engine based on distributed acquisition strategy [D]. Harbin: Northeast Forestry University, 2017.)

[28] 孟繁疆, 姬祥, 袁琦, 等. 农产品价格主题搜索引擎的研究与实现 [J]. 东北农业大学学报, 2016, 47 (9): 64-71. (Meng Fanjiang, Ji Xiang, Yuan Qi, *et al.* Research and implementation of agricultural product price subject search engine [J]. Journal of Northeast Agricultural University, 2016, 47 (9): 64-71.)

[29] 尹曼. 医疗器械垂直搜索引擎的设计与实现 [D]. 重庆：重庆大学. 2017 年. (Yin Man. Design and implementation of medical equipment vertical search engine [D]. Chongqing: Chongqing University. 2017)

[30] 李学博. 基于 Hadoop 的中医药 Web 信息资源评价体系研究 [D]. 济南：山东医药大学. 2016. (Li Xuebo. Research on Chinese medicine web information resource evaluation system based on hadoop [D]. Jinan: Shandong Medical University, 2016.)

[31] 刘灿, 任剑宇, 李伟, 等. 面向个性化推荐的教育新闻爬取及展示系统 [J]. 软件工程, 2018, 21 (2): 34-40. (Liu Can, Ren Jianyu, Li Wei, *et al.* Education news crawling and display system for personalized recommendations [J]. Software Engineering, 2018, 21 (2): 34-40.)

[32] Hu Rong, Feng Zhongke, Jiang Junzhiwei. Web crawler of atmosphere and weather data based on MEAN stack with CheerIO [J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47 (6): 275-282.

[33] 李翔宇. 生物安全领域本体建模与知识平台开发 [D]. 天津：天津大学, 2016. (Li Xiangyu. Ontology modeling and knowledge platform development in biosafety field [D]. Tianjin: Tianjin University, 2016.)

[34] 关卫国. 面向食品接触材料安全的爬虫系统设计与实现 [D]. 上海：东华大学. 2017. (Guan Weiguo. Design and implementation of reptile system for food contact material safety [D]. Shanghai: Donghua University, 2017.)

[35] Ding Shenchun, Gong Silan, Zhou Wenjie, *et al.* Research on network public opinion real-time monitoring of the south china sea issue based on knowledge base and focused crawler [J]. Journal of Intelligence, 2016, 35 (5): 34-37.